

C  
E  
N  
T  
R  
E  
B  
I  
O  
I  
N  
F  
O  
R  
M  
A  
T  
I  
C  
S

## Computational Genomics and Proteomics 2008

### Lecture 6: DNA as a Turing Machine: Gene prediction in prokaryotes (bacteria)

ibivu.nl

## Biology – Bioinformatics - Evolution

“Nothing in biology makes sense except in the light of evolution” (Theodosius Dobzhansky (1900-1975))

↓

“Nothing in bioinformatics makes sense except in the light of biology”

“Nothing in Bioinformatics makes sense except in the light of evolution”

## Functional Genomics

*From gene to function*

We need to know the genes to be able to build the cellular networks and understand how an organism works

Gene prediction (delineating where the genes are in a genome) is a crucial first step

Genome  
Expressome  
Proteome  
Metabolome

## Turing machine

A Turing machine consists of a

- **TAPE** (of (semi-)infinite length) and a
- **HEAD** (reading/writing the tape) combined with a **TABLE** (instruction set)

## Turing machine

- mathematically models a machine that mechanically operates on a tape on which symbols are written, which it can read and write one at a time using a tape head;
- operation is fully determined by a finite set of elementary instructions, such as "in state 42, if the symbol you see is a '0', write a '1'; if you see a '1', shift to the right, and change into state 17; in state 17, if you see a '0', write a '1' and change to state 6;" et cetera.
- In the original article ("On computable numbers, with an application to the Entscheidungsproblem"), Turing imagines not a mechanical machine, but a person, whom he calls the "computer", who executes these deterministic, mechanical rules slavishly (or as Turing puts it, "in a desultory manner").

## Alan Turing: the person

**Alan Mathison Turing**, OBE, FRS (23 June 1912 – 7 June 1954) was an English mathematician, logician and cryptographer.

Turing is often considered to be the father of modern computer science. He provided an influential formalisation of the concept of the algorithm and computation with the **Turing machine**. With the **Turing test** meanwhile, he made a significant and characteristically provocative contribution to the debate regarding artificial intelligence: whether it will ever be possible to say that a machine is conscious and can think. He later worked at the National Physical Laboratory, creating one of the first designs for a stored-program computer, the ACE, although it was never actually built in its full form. In 1948, he moved to the University of Manchester to work on the Manchester Mark I, then emerging as one of the world's earliest true computers. During the Second World War Turing worked at Bletchley Park, the UK's code breaking centre, and was for a time head of Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including the method of the bombe an electromechanical machine that could find settings for the **Enigma machine**.

Turing also developed concepts in theoretical **biology**.

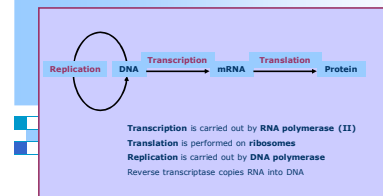
After WWII, Turing was prosecuted for homosexuality and committed suicide in 1954.



## The codon table

		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } Ser UCC } UCA } UCG }	UAU } Tyr UAC } UAA } Stop UAG } Stop	UGU } Cys UGC } UGA } Stop UGG } Trp	U	C A G
	C	CUU } Leu CUC } CUA } CUG }	CCU } Pro CCC } CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } Arg CGC } CGA } CGG }	U C A G	
	A	AUU } Ile AUC } AUA } Met AUG }	ACU } Thr ACC } ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	
	G	GUU } Val GUC } GUA } GUG }	GCU } Ala GCC } GCA } GCG }	GAU } Asp GAC } GAA } glu GAG }	GGU } Gly GGC } GGA } GGG }	U C A G	

## Central Dogma of Molecular Biology



Transcription + Translation = Expression

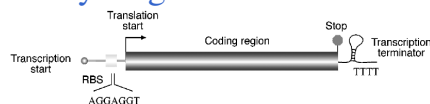
## But DNA can also be transcribed into non-coding RNA ...

- q tRNA (transfer): transfer of amino acids to the ribosome during protein synthesis.
- q rRNA (ribosomal): essential component of the ribosomes (complex with rProteins).
- q snRNA (small nuclear): mainly involved in RNA-splicing (removal of introns). *snRNPs*.
- q snoRNA (small nucleolar): involved in chemical modifications of ribosomal RNAs and other RNA genes. *snoRNPs*.
- q SRP RNA (signal recognition particle): form RNA-protein complex involved in mRNA secretion.
- q Further: microRNA, eRNA, gRNA, tmRNA etc.

## Prokaryote gene

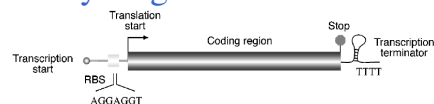
- Prokaryotes, which include bacteria and Archaea, have relatively small genomes with sizes ranging from 0.5 to 10Mbp (1Mbp=10<sup>6</sup> bp).
- The gene density in the genomes is high, with more than 90% of a genome sequence containing coding sequence.
- There are very few repetitive sequences. Each prokaryotic gene is composed of a single contiguous stretch of ORF coding for a single protein or RNA with no interruptions within a gene (no splicing).

## Prokaryote gene



- The majority of genes have a start codon ATG (or AUG in mRNA) coding for methionine. Occasionally, GTG and TTG are used as alternative start codons, but methionine is still the actual amino acid inserted at the first position.
- Because there may be multiple ATG, GTG, or TGT codons in a frame, the presence of these codons at the beginning of the frame does not necessarily give a clear indication of the translation initiation site.
- To help gene identification, other features associated with translation are used: One such feature is the **ribosomal binding site (RBS)**, also called the **Shine-Delgarno sequence**, a stretch of purine-rich sequence complementary to 16S rRNA in the ribosome. It is located immediately downstream of the transcription initiation site and slightly upstream of the translation start codon. In many bacteria, it has a consensus motif of AGGAGGT. Identification of the ribosome binding site can help locate the start codon.

## Prokaryote gene



- At the end of the protein coding region is a stop codon that causes translation to stop. There are three possible stop codons, identification of which is straightforward.
- Many prokaryotic genes are transcribed together as one operon. The end of the operon is characterized by a transcription termination signal called  *$\rho$ -independent terminator*. The terminator sequence has a distinct stem-loop secondary structure followed by a string of Ts. Identification of the terminator site, in conjunction with promoter site identification, can sometimes help in gene prediction.

## Prokaryote gene prediction how to predict an ORF by hand

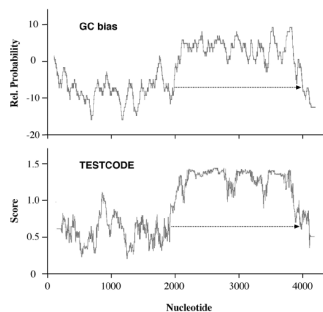
- Perform conceptual translation in all six possible frames, three frames forward and three frames reverse. Because a stop codon occurs in about every twenty codons by chance in a noncoding region, a frame longer than 30 codons without interruption by stop codons is suggestive of a gene coding region (threshold is normally set even higher at 50 or 60 codons).
- The putative frame is further manually confirmed by the presence of other signals such as a start codon and Shine–Delgarno sequence.
- Furthermore, the putative ORF can be translated into a protein sequence, which is then used to search against a protein database. Detection of homologs from this search is probably the strongest indicator of a protein-coding frame.

## Prokaryote gene prediction how to predict an ORF computationally

### Examine nonrandomness of nucleotide distribution

- **GC bias:** the third position of a codon has a preference to use G or C over A or T in a coding sequence. By plotting the GC composition at this position, regions with values significantly above the random level can be identified, which are indicative of the presence of ORFs. In practice, the statistical patterns are computed for all six possible frames.
- The method TESTCODE (implemented in the commercial GCG package) exploits the fact that the third codon nucleotides in a coding region tend to repeat themselves. By plotting the repeating patterns of the nucleotides at this position, coding and noncoding regions can be differentiated. The results of the two methods are often consistent.
- These two early methods are often used in conjunction to confirm the results of each other.

## Prokaryote gene prediction



Coding frame detection of a bacterial gene using either the GC bias or the TESTCODE method. Both result in similar identification of a reading frame (*dashed arrows*).

## Prokaryote gene prediction

### Gene Prediction Using Markov Models and Hidden Markov Models

Markov models and HMMs can be very helpful in providing finer statistical description of a gene.

A Markov model describes the probability of the distribution of nucleotides in a DNA sequence, in which the conditional probability of a particular sequence position depends on  $k$  previous positions.  $k$  is the order of a Markov model.

- A zero-order Markov model assumes each base occurs independently with a given probability. This is often the case for noncoding sequences.
- A first-order Markov model assumes that the occurrence of a base depends on the base preceding it.
- A second-order model looks at the preceding two bases to determine which base follows, which is more characteristic of codons in a coding sequence.

## Prokaryote gene prediction

### Gene Prediction Using Markov Models and Hidden Markov Models

- The use of Markov models in gene finding exploits the fact that oligonucleotide distributions in the coding regions are different from those for the noncoding regions.
- These can be represented with various orders of Markov models. Since a fixed-order Markov chain describes the probability of a particular nucleotide that depends on previous  $k$  nucleotides, the longer the oligomer unit, the more nonrandomness can be described for the coding region. Therefore, the higher the order of a Markov model, the more accurately it can predict a gene.

## Prokaryote gene prediction

### Gene Prediction using Hidden Markov Models

- Because a protein-encoding gene is composed of nucleotides in triplets as codons, more effective Markov models are built in sets of three nucleotides, describing nonrandom distributions of trimers or hexamers, and so on.
- The parameters of a Markov model have to be trained using a set of sequences with known gene locations. Once the parameters of the model are established, it can be used to compute the nonrandom distributions of trimers or hexamers in a new sequence to find regions that are compatible with the statistical profiles in the learning set.

## Prokaryote gene prediction

### Gene Prediction using Hidden Markov Models

- Statistical analyses have shown that pairs of codons (or amino acids at the protein level) tend to correlate. The frequency of six unique nucleotides appearing together in a coding region  $i$  is much higher than by random chance. Therefore, a fifth-order Markov model, which calculates the probability of hexamer bases, can detect nucleotide correlations found in coding regions more accurately and is in fact most often used.
- A potential problem of using a fifth-order Markov chain is that if there are not enough hexamers, which happens in short gene sequences, the method's efficacy may be limited.

## Prokaryote gene prediction

### Gene Prediction using Hidden Markov Models

- To cope with this limitation, a variable-length Markov model, called an *interpolated Markov model (IMM)*, has been developed.
- The IMM method samples the largest number of sequence patterns with  $k$  ranging from 1 to 8 (dimers to ninemers) and uses a weighting scheme, placing less weight on rare  $k$ -mers and more weight on more frequent  $k$ -mers.
- The probability of the final model is the sum of probabilities of all weighted  $k$ -mers. In other words, this method has more flexibility in using Markov models depending on the amount of data available. Higher-order models are used when there is a sufficient amount of data and lower-order models are used when the amount of data is smaller.

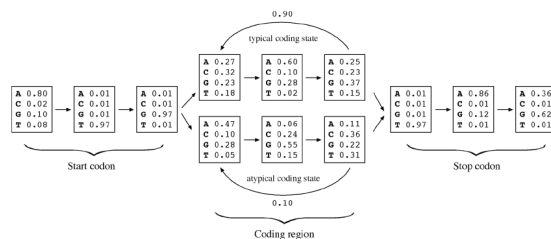
## Prokaryote gene prediction

### Gene Prediction using Hidden Markov Models

- It has been shown that the gene content and length distribution of prokaryotic genes can be either typical or atypical. Typical genes are in the range of 100 to 500 amino acids with a nucleotide distribution typical of the organism. Atypical genes are shorter or longer with different nucleotide statistics. These genes tend to escape detection using the typical gene model. This means that, to make the algorithm capable of fully describing all genes in a genome, more than one Markov model is needed.
- To combine different Markov models that represent typical and atypical nucleotide distributions leads to a HMM prediction algorithm.

## Prokaryote gene prediction

### Gene Prediction using Hidden Markov Models



A simplified second-order HMM for prokaryotic gene prediction that includes a statistical model for start codons, stop codons, and the rest of the codons in a gene sequence represented by a typical model and an atypical model.

## HMM prokaryote gene prediction methods

The following describes a number of HMM/IMM-based gene finding programs for prokaryotic organisms.

- **GeneMark** (<http://opal.biology.gatech.edu/GeneMark/>) is a suite of gene prediction programs based on the fifth-order HMMs.
  - The main program—GeneMark.hmm—is trained on a number of complete microbial genomes. If the sequence to be predicted is from a nonlisted organism, the most closely related organism can be chosen as the basis for computation.
- Another option for predicting genes from a new organism is to use a self-trained program **GeneMarkS** as long as the user can provide at least 100 kbp of sequence on which to train the model.
  - If the query sequence is shorter than 100 kbp, a GeneMark heuristic program can be used with some loss of accuracy.
- In addition to predicting prokaryotic genes, GeneMark also has a variant for eukaryotic gene prediction using HMM.

## HMM prokaryote gene prediction methods

- **Glimmer** (Gene Locator and Interpolated Markov Modeler, [www.tigr.org/softlab/glimmer/glimmer.html](http://www.tigr.org/softlab/glimmer/glimmer.html)) is a UNIX program from TIGR that uses the IMM algorithm to predict potential coding regions. The computation consists of two steps, namely model building and gene prediction. The model building involves training by the input sequence, which optimizes the parameters of the model. In an actual gene prediction, the overlapping frames are “flagged” to alert the user for further inspection.
- Glimmer also has a variant, GlimmerM, for eukaryotic gene prediction.

## HMM prokaryote gene prediction methods

- **FGENESB**

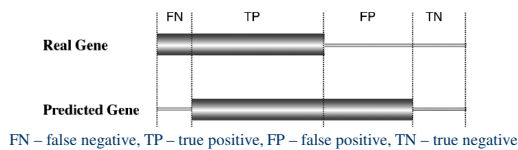
([www.softberry.com/berry.phtml?topic=gfindb](http://www.softberry.com/berry.phtml?topic=gfindb)) is a web-based program that is also based on fifth-order HMMs for detecting coding regions. The program is specifically trained for bacterial sequences. It uses the Viterbi algorithm to find an optimal match for the query sequence with the intrinsic model (the Viterbi algorithm can be implemented using Dynamic Programming). A linear discriminant analysis (LDA) is then used to further distinguish coding signals from noncoding signals.

## HMM prokaryote gene prediction methods

- These programs on the earlier slides have been shown to be reasonably successful in finding genes in a genome. The common problem is imprecise prediction of translation initiation sites because of inefficient identification of ribosomal binding sites. This problem can be remedied by identifying the ribosomal binding site associated with a start codon. A number of algorithms have been developed solely for this purpose.
- RBSfinder is one such algorithm. RBSfinder (<ftp://ftp.tigr.org/pub/software/RBSfinder/>) is a UNIX program that uses the prediction output from Glimmer and searches for the Shine–Delgarno sequences in the vicinity of predicted start sites. If a high-scoring site is found by the intrinsic probabilistic model, a start codon is confirmed; otherwise the program moves to other putative translation start sites and repeats the process.

## Prokaryote gene prediction

### - prediction success



**Sensitivity:**  $S_n = TP / (TP + FN)$   
**Specificity:**  $S_p = TP / (TP + FP)$  -- generally called *Positive Predictive Value (PPV)*

**Matthews correlation:**  $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

## Prokaryote gene prediction

### Performance analysis of the Glimmer program for gene Prediction of three genomes

Species	GC (%)	FN	FP	Sensitivity	Specificity
<i>Campylobacter jejuni</i>	30.5	10	19	99.3	98.7
<i>Haemophilus influenzae</i>	38.2	3	54	99.8	96.1
<i>Helicobacter pylori</i>	38.9	6	39	99.5	97.2

Note: The data sets were from three bacterial genomes (Aggarwal and Ramaswamy, 2002).  
 Abbreviations: FN, false negative; FP, false positive.

The methods overview in this lecture has been largely based on Jin Xiong's *Essential Bioinformatics* -- Chapter 8.

END